Auditing Indian Elections

Vishal Mohanty[1], Nicholas Akinyokun[2], Andrew Conway[3], Chris Culnane[2], Philip B. Stark[4] and Vanessa Teague[2]

[1]Department of Computer Science and Engineering, Indian Institute of Technology, Madras, E-mail: `cs15b039@smail.iitm.ac.in`, Tel. `+91 9840298064`, This work was done while visiting the University of Melbourne
[2] School of Computing and Information Systems, University of Melbourne, {`christopher.culnane, oakinyokun, pstuckey, vjteague`}`@unimelb.edu.au`
[3]aindaud@greatcactus.org
[4]Department of Statistics, University of California, Berkeley, `stark@stat.berkeley.edu`

January 28, 2019

Abstract

Indian Electronic Voting Machines (EVMs) will be fitted with printers that produce Voter-Verifiable Paper Audit Trails (VVPATs) in time for the 2019 general election. VVPATs provide evidence that each vote was recorded as the voter intended, without having to trust the perfection or security of the EVMs. However, confidence in election results requires more: VVPATs must be preserved inviolate and then actually used to check the reported election result in a trustworthy way that the public can verify. A full manual tally from the VVPATs could be prohibitively expensive and time-consuming; moreover, it is difficult for the public to determine whether a full hand count was conducted accurately. We show how *Risk-Limiting Audits* (RLAs) could provide high confidence in Indian election results. Compared to full hand recounts, RLAs typically require manually inspecting far fewer VVPATs when the outcome is correct, and are much easier for the electorate to observe in adequate detail to determine whether the result is trustworthy.

We discuss how two RLA strategies, *ballot-level comparison* and *ballot polling*, can be applied to General Elections in India. Our main result is a novel method for combining RLAs in constituencies to obtain an RLA of the overall parliamentary election result.

1   Introduction

Since Electronic Voting Machines (EVMs) were introduced in India in the 1999 elections, there have been questions about their transparency and trustworthiness; a number of security vulnerabilities have been documented (Wolchok et al. (2010)). In 2013, the Indian Supreme Court ruled that all EVMs in Indian General Elections must be equipped with printers providing Voter-Verifiable Paper Audit Trails (VVPATs, Supreme Court of India (2013)). The Election Commission of India has introduced VVPAT-equipped EVMs in several constituencies and has promised that all EVMs used in the upcoming 2019 General Election will have VVPAT printers.

VVPATs allow each voter to verify that his or her intended selections are correctly printed on a paper record, which is collected in a separate container called the *VVPAT box*. VVPATs provide a way to check and correct election results, for instance, if there is a legal demand by a candidate, or for routine checks of election tabulation accuracy—audits. VVPATs could be manually recounted to check the electronic results, but that is labor-intensive and time-consuming. We show how auditing a random sample of VVPAT records can justify confidence in election results without a full manual tally. Auditing a VVPAT means manually inspecting the paper record to see the voter preferences it shows. Different auditing strategies—for instance, ballot-level comparison audits and ballot-polling audits—use that information differently, as described below.

The Election Commission (EC) of India is taking steps to increase the transparency of the Indian electoral system. In a recent report,[1] the EC decided to tally the paper trail slips and compare them with the electronic result provided by the EVMs in 5% of the booths in each Assembly seat district, selected randomly. This effort, while well-intentioned, does not suffice to give strong evidence that election results are correct. In this paper, we show rigorous ways of attaining well-defined confidence levels.

Suitable post-election audits may justify confidence of voters, candidates, and parties that election results are correct. One type of post-election audit is a Risk-Limiting Audit (RLA), which either develops strong statistical evidence that the reported outcome is correct, or corrects the results (by conducting a full manual tally of a reliable paper trail). Here, "outcome" means the set of reported winners of the contests, not the exact vote tallies. To ensure that the tallies are correct to the last vote is prohibitively expensive, if not impossible; conversely, to ensure that the reported winners really won seems like the lowest reasonable standard for accuracy.

Before a RLA commences, the *risk limit* (denoted α) must be chosen; ideally, it is set in legislation or regulation, so that auditors cannot manipulate the level of scrutiny a contest gets by adjusting the risk limit. The risk limit is the maximum probability that the audit will fail to correct the reported election outcome, on the assumption that the reported outcome is wrong. The risk limit is a worst-case probability that makes no assumption about *why* the outcome is wrong, *e.g.*, it could be because of accidental error, procedural lapse, bugs, misconfiguration, or malicious hacking by a strategic adversary who knows how the audit will be conducted. RLAs assume that the paper ballots reflect the correct outcome, *i.e.* that a full manual tally of the paper trail would show who really won. A RLA of an unreliable paper

---

[1] http://indianexpress.com/article/india/ec-to-tally-paper-trail-slips-with-evms-in-5-pc-booths-in-ea

trail is "security theater." Hence, there need to be procedures (called *compliance audits* by Benaloh et al. (2011); Lindeman and Stark (2012); Stark and Wagner (2012); Stark (2018)) to ensure that the paper trail is complete and intact before the RLA begins.

This paper shows how two types of RLAs can be used with Indian elections: transitive ballot-level comparison audits and ballot-polling audits. Ballot-level comparison audits require a "commitment" to the interpretation of each ballot in a way that allows that interpretation to be compared to a human reading of voter intent directly from the paper ballot. Ballot-polling audits do not require knowing how individual ballots were interpreted. Ballot-level comparison audits are more efficient in the sense that they generally involve inspecting fewer ballots to attain the same risk limit when the reported outcome is correct. However, they require more setup. As discussed in Section 3, they may require a voting system that can export its interpretation of individual ballots in a way that can be matched to the corresponding paper, or may require sorting the physical ballots or VVPATs before the audit, according to the votes they (reportedly) show.

Our main contribution is to develop RLAs for a new social choice function—Indian parliamentary majorities—with procedures suited to the logistics of Indian elections. To verify the overall election outcome we need to verify that the party/coalition reported to have been elected to form the government actually won. That generally requires less auditing than confirming the winner in every constituency. The method we develop splits the responsibility of the auditing among various constituencies in a way that the combined result gives higher confidence in the correctness of the overall parliamentary outcome than each constituency would have in its results alone. This procedure is discussed in Section 4. Our methods apply to any parliamentary democracy, but the computations are particularly simple when all constituencies have equal weight.

2   Background

2.1   Indian Elections

Indian General elections are held Quinquennially to elect the Lok Sabha (Lower House of the Parliament). The country is divided into 543 constituencies, each represented by one person elected to the Lok Sabha. Elections at the constituency level are *plurality* contests: the person who gets the most votes wins. Candidates at the constituency level typically belong to some political party, but can be unaffiliated with any party. At the parliamentary level, the party that gets the *majority* of the seats forms the government. If no party has a majority, parties may form coalitions to attain a majority. Coalitions can be formed before or after elections, although before is more common. Elections are conducted in phases spread over a month. Each phase consists of single-day elections in a subset of constituencies, typically grouped by geography.

2.2   Related work on election auditing

RLAs are procedures that guarantee a minimum chance of conducting a full manual tally of the voter-verifiable records when the result of that tally would belie the reported outcome.

They amount to a statistical test of the *null hypothesis* that the election outcome is wrong, at significance level α, the *risk limit*, chosen in advance. A RLA continues to examine more ballots until the null hypothesis is rejected at significance level α, allowing one to conclude with known confidence that the outcome is in fact correct—or until there has been a complete manual tally to set the record straight. The risk limit is the largest chance that the audit will *not* require a full manual tally of the paper records if the electoral outcome according that tally would differ from the reported electoral outcome.

RLAs were introduced by Stark (2008), but were not so named until Stark (2009). The first RLAs were conducted in California in 2008 (Hall et al. (2009)). RLAs have been conducted in California, Colorado, Indiana, Michigan, New Jersey, Ohio, Virginia, and Denmark. RLAs have been developed for a variety of social choice functions and for variety of sampling strategies (unstratified sampling of individual ballots or batches, with or without replacement; stratified sampling with and without replacement, Bernoulli sampling) and auditing strategies (*batch-level comparisons*, *ballot-level comparisons*, and *ballot polling*).

Ballot-polling audits (Lindeman, Stark, and Yates (2012); Lindeman and Stark (2012); Ottoboni, Bernhard, et al. (2019)) do not require knowing how the system interpreted individual ballots nor how it tallied the votes on subsets of ballots. They directly check whether the reported winner(s) received more votes than the reported loser(s) by sampling and manually interpreting individual ballots. To draw a random sample of ballots typically involves a *ballot manifest*, which describes how the physical ballots are organized: the number of bundles, the labels of the bundles, and the number of ballots in each bundle. (However, see Ottoboni, Bernhard, et al. (2019).)

The BRAVO ballot-polling method (Lindeman, Stark, and Yates (2012)) uses Wald's sequential probability ratio test (Wald (1945)) to test a collection of null hypotheses, namely, that any loser in fact tied or beat any winner. For each (winner, loser) pair, the audit tests the hypothesis that the loser got as many or more votes than the winner. The audit stops (short of a full hand count) only when there is sufficiently strong evidence that every winner beat every loser.

Comparison audits involve manually checking the voting system's interpretation of the votes on physically identifiable subsets of ballots. They require the voting system to export vote tallies for physically identifiable subsets of ballots, so that the votes on those ballots can be tallied by hand and compared to the voting system's tallies. They also require checking that the reported subtotals yield the reported contest results, and that the subtotals account for all ballots cast in the contest. They generally also require ballot manifests.

A comparison audit that checks the voting system's interpretation of individual ballots is a *ballot-level comparison* RLA. Ballot-level comparison RLAs are more efficient than batch-level comparison RLAs and ballot-polling RLAs in that they generally require examining fewer ballots when the reported outcome is correct. However, they have higher set-up costs and require more data export from the voting system: they need a *cast vote record* or *CVR* for each physical ballot, a way to locate the CVR for each physical ballot, and *vice versa*. (A CVR is the voting system's interpretation of voter intent for a given ballot.) Relying on more general results in Stark (2009) for batch-level comparison audits, Stark (2010) developed a sequential ballot-level comparison RLA method that results in particularly simple calculations.

Transitive audits (Calandrino, J. Halderman, and E.W. Felten (2007); Lindeman and Stark (2012)) involve auditing an unofficial system that is easier to audit than the official

system. If the two systems agree who won, an audit that provides strong evidence that the unofficial system found the correct winner(s) transitively provides strong evidence that the official system did also; and if the audit of the unofficial system leads to a full manual tally, the outcome of that tally can be used to correct the official result. A transitive audit does not confirm that the official system tallied votes correctly; indeed, the two systems might disagree about the interpretation of every ballot, but still agree who won.

Indian EVMs do not create CVRs, but—if the VVPATs are organized appropriately— they can still be audited using a transitive ballot-level comparison audit. CVRs can be constructed for EVMs by sorting the VVPATs into bundles that (purportedly) show the same voter preferences, counting the number of VVPATs in each batch, and labeling each bundle with the number of ballots and the voter preferences it purports to contain. A report of the bundle labels, the number of VVPATs in each bundle, and the reported voter preference for the bundle amounts to a CVR for every VVPAT. Such a report in effect combines a *ballot manifest* (Lindeman and Stark (2012)) and a commitment to a cast vote record for every ballot, implied by the label of the bundle the ballot is in. We shall call such a report a *preference manifest.*

If ballots are sufficiently simple—*e.g.*, if each contains only one contest, as in India— sorting ballots by voter preference can be practical. Indeed, this is how ballots are tallied manually in Denmark: on election night, ballots are sorted within polling places according to the voter's party preference. The following day, ballots are sorted further according to the voter's candidate preference, to produce homogeneous bundles of ballots, each labeled with the number of ballots and the voters' preference.

Such sorting-based CVRs were the basis of a RLA in Denmark (Schürmann (2016)). The sorting might be manual, as it is in Denmark, but it could be automated partly or entirely. (Sorting may also increase vote anonymity by breaking any link between voter and ballot.) When the official tallying process itself is based on creating and counting the homogeneous bundles, as it is in Denmark, the audit is a direct audit of the voting system. If the sorting is conducted independently of the tabulation, as it would be if India were to sort the paper ballots to produce a preference manifest, the resulting audit is a transitive audit.

The first step of a ballot-level comparison RLA is to verify that the CVRs are consistent with the reported results: that applying the social choice function to the vote subtotals implied by the sizes of the bundles and the votes they purport to contain produces the same set of winners. If the preference manifest does not produce the same set of winners as reported, the audit should not continue: there is a serious problem. The audit should also check that the number of CVRs for each contest does not exceed the number of ballots cast in the contest, which should be determined without reliance on the voting system (Bañuelos and Stark (2012)). If the preference manifest passes these checks, the audit can begin to select ballots at random to check the accuracy of the CVRs implied by the preference manifest against a manual reading of voter intent from each paper ballot.

Kroll, A. Halderman, and E. Felten (2014) present a method for reducing the workload in auditing multi-level elections, inspired by the US Electoral College. They show that to achieve an overall confidence that a party or coalition secured the majority of seats, the individual constituencies can sometimes be audited to lower confidence levels. They provide a constraint optimization program describing the set of feasible solutions (*i.e.* those that constitute a sufficient audit) and a number of methods for finding the optimal solution. In

India's electoral system, as in many other parliamentary democracies, every constituency has equal weight.

3   Auditing individual constituencies using extant methods

This section discusses how existing methods for RLAs apply to Indian elections. We consider auditing individual constituencies rather than the entire election; Section 4 shows how to combine audits of constituencies to audit an entire contest.

India's voting system currently does not support ballot-level comparison audits, but, as described above, if procedures were added to sort the paper ballots and to report a preference manifest, transitive ballot-level comparison audits would be possible. Because ballots in India are simple—a single selection in a single contest—such sorting is feasible.

Ballot-polling audits could be used in India without requiring sorting the ballots or modifying the voting system, if ballot manifests were available (see section 3.2). The calculations for BRAVO (Lindeman, Stark, and Yates (2012)) and the ballot-polling method in Lindeman and Stark (2012) are simple enough to do with a pencil and paper or hand calculator, and have open-source online tools by Stark: `https://www.stat.berkeley.edu/~stark/Vote/ballotPollTools.ht`

When the election outcome is correct, ballot-level comparison audits generally require inspecting fewer ballots than ballot-polling audits. (Because they are RLAs, when the outcome is incorrect, both methods have a large chance of requiring a full manual tally.) The advantage grows as the margin shrinks: as a rule of thumb, workload increases inversely with the reported margin for ballot-level comparison audits, and increases inversely with the square of the actual margin for ballot-polling audits. However, preparing for a ballot-level comparison audit is harder, because it requires CVRs linked to the corresponding physical ballots. While they may require inspecting more ballots, the simplicity of ballot-polling audits may offset the work of examining more paper, unless the margin is very small.

3.1   Transitive Ballot-Level Comparison RLA

Ballot-level comparison RLAs were introduced by Stark (2010) who provides online tools at `https://www.stat.berkeley.edu/~stark/Vote/auditTools.htm`; see also Lindeman and Stark (2012). Ballot-level comparison audits require a way to find the CVR corresponding to each paper ballot, and vice versa. The EVMs currently used in India do not provide CVRs at all.

However, sorting ballots into groups according to the vote (if any) that they are reported to show in effect provides a CVR for each ballot through a preference manifest that lists the bundles of ballots, the number of ballots in each bundle, and the (single) preference that every ballot in the bundle is supposed to show. This approach was used by Schürmann (2016) to audit an election in Denmark.

In Denmark, ballots are manually sorted into bundles with homogeneous voter intent, but sorting could be automated with relatively simple equipment, possibly something similar to the system used in South Korea. Whether it is worth the effort to sort the ballots depends on the margin in the contest: if the margin is wide, it will be less expensive to use ballot polling,

but if the margin is very narrow, the cost of sorting—whether manual or automated—may reduce the sample size required to confirm the outcome by orders of magnitude.

3.1.1   Classifying CVR errors

Stark (2009) reviews a number of methods to test the hypothesis that any loser received more votes than any winner by comparing hand counts of votes in randomly selected batches of ballots to the machine counts of the votes on the same ballots. The methods apply to arbitrarily small batches, including batches consisting of a single ballot; that is, to ballot-level comparison audits. Stark (2010) elaborated on one of those methods, which relies on the Kaplan-Markov inequality. By introducing a taxonomy of discrepancies, the arithmetic can be simplified to the point that a pencil and paper suffice, while rigorously maintaining the risk limit. That "super-simple simultaneous single-ballot" method was further simplified by Lindeman and Stark (2012), and is the basis of pilot audits in Denmark, California, Colorado, Indiana, Michigan, New Jersey, and Virginia, and of the statutory risk-limiting audits in most Colorado counties.

Stark and Teague (2014) presented a ballot-level comparison RLA method based on the Kaplan-Wald inequality, which has some advantages over the Kaplan-Markov inequality. In this paper, we use the Kaplan-Markov method as simplified by Lindeman and Stark (2012), because it has been used more widely. We shall refer to it as the *LSKM method*. It is straightforward to modify the procedures below to use the method of Stark and Teague (2014) instead.

The LSKM method is *sequential*: it involves examining more and more ballots selected at random until either there is strong evidence that the reported winners really won, or until there has been a full hand count and the correct outcome is known. Conceptually, after examining one or more ballots, one calculates a sequentially valid[2] *P*-value of the hypothesis that the outcome is wrong. If that *P*-value is less than or equal to the risk limit, the audit stops; otherwise, more ballots are audited and the sequential *P*-value is updated. The method presented below to check the overall electoral outcome involves combining the *P*-values for individual constituencies.

If the audit does lead to a full hand tally in a constituency, the reported results are replaced by the results according to that full hand tabulation. Election officials may elect to terminate the audit and conduct a full hand count at any time, for instance, if they estimate that the cost of additional sampling will exceed the cost of a full manual tally.

The LSKM method involves classifying discrepancies between the CVR and a manual reading of voter intent from the paper ballot:

- If correcting the CVR would reduce the margin between any (reported) winner and any (reported) loser by two votes, the discrepancy is a *2-vote overstatement* (the number of 2-vote overstatements is denoted o_2).

---

[2] *Sequentially valid* means that the chance that the infimum of the *P*-value over all sample sizes is less than or equal to p is less than or equal to p if the null hypothesis is true. In contrast, standard hypothesis tests are designed for sample sizes that are fixed ahead of time: expanding the sample and re-calculating the *P*-value for such tests generally produces type I error rates far larger than the nominal significance level, because it does not account for multiplicity.

- If not, but if correcting the CVR would reduce the margin between any winner and any loser by one vote, the discrepancy is a *1-vote overstatement* (the number of 1-vote overstatements is denoted o_1).

- If not, but if correcting the CVR would not increase the margin between every winner and every loser, the discrepancy is a *neutral error*. (Neutral errors do not enter the stopping rule explicitly.)

- If not, but if correcting the CVR would increase the margin between every winner and every loser by at least one vote, and increase the margin between some winner and some loser by exactly one vote, the discrepancy is a *1-vote understatement* (the number of 1-vote understatements is denoted u_1).

- If correcting the CVR would increase the margin between every winner and every loser by two votes, it is a *2-vote understatement* (the number of 2-vote understatements is denoted u_2).

Two-vote overstatements should be rare if the voting system is working correctly: they involve mistaking a vote for a loser as a vote for a winner. Two-vote understatements should be even rarer—and are typically mathematically impossible. For instance, in a plurality, vote-for-one contest with three or more candidates, two-vote understatements are impossible, because they would require having mistaken a valid vote for the winner as a valid vote for every losing candidate.

We assume that there is a trustworthy upper bound on the total number of ballots cast, for instance, from pollbooks or from information about the number of eligible voters. A preliminary check should ensure that the preference manifest does not list more ballots than that upper bound: if there are more ballots listed than can exist, there is a serious problem that the audit cannot address by itself.[3]

In the sorted-ballot method described above,

- a *2-vote overstatement* occurs if we find a vote for a reported loser in the reported winner's pile;

- a *1-vote overstatement* occurs if we find a vote for a different reported loser in a reported loser's pile;

- *neutral errors* don't occur;[4]

- a *1-vote understatement* occurs when there are at least three candidates and we find a vote for the reported winner in a reported loser's pile;

- a *2-vote understatement* occurs only when there are exactly two candidates, and we find a vote for the reported winner in the reported loser's pile;

---

[3]Prof Sandeep Shukla of IIT Kanpur has pointed out that the current Indian VVPAT design does not protect against the EVM adding electronic votes and corresponding VVPATs when the voter is not looking, because there is no publicly observable mechanism to ensure that at most one VVPAT is inserted into the box per voter. This needs to be addressed by improving the physical design in a way that is out of the scope of this paper.

- if a pile turns out to be *smaller* than reported, the discrepancy can be addressed using the "phantom to zombie" approach of Bañuelos and Stark (2012).

- if a pile turns out to be *larger* than reported, then some other pile must be *smaller* than reported, and the "phantom to zombie" approach of Bañuelos and Stark (2012) will still ensure that the risk is controlled conservatively.

There are sharper ways to treat discrepancies than to use these categories (in particular, keeping track of *which* margins are affected by each discrepancy can reduce the number of ballots the audit inspects; see Stark (2010)). However, the bookkeeping is more complex. Categorizing discrepancies this way makes the calculations simple enough to do with a pencil and paper (aside from calculating 5 constants involving logarithms, which can be done once and for all and verified by anyone).

3.1.2   Calculations

Let n denote the current sample size and α the risk limit. Fix $\gamma \geq 1$. The LSKM method stops auditing (and concludes that the reported winners really won) if

$$n \geqslant \frac{2\gamma}{\mu} \left(o_1 \log \left(\frac{1}{1 - \frac{1}{2\gamma}} \right) + o_2 \log \left(\frac{1}{1 - \frac{1}{\gamma}} \right) - u_1 \log \left(1 + \frac{1}{2\gamma} \right) - u_2 \log \left(1 + \frac{1}{\gamma} \right) - \log(\alpha) \right).$$

(3.1)

In this expression, μ is the *diluted margin*, the smallest difference in votes between any winner and any loser, divided by the total number of ballots in the population from which the sample is to be drawn, including ballots with invalid votes. The constant $\gamma \geq 1$ is the *error inflation factor*, which controls the operating characteristics of the LSKM method when errors are observed: the larger γ, the fewer additional ballots need to be audited if a 2-vote overstatement is observed, but the smaller γ is, the fewer ballots need to be audited if no 2-vote overstatements are observed. Because two-vote overstatements should be rare, taking γ slightly larger than 1 should suffice. For γ exactly equal to 1, then if the audit finds even one 2-vote overstatement, the audit will not terminate without a full hand count. Lindeman and Stark (2012) suggest using $\gamma = 1.03905$, which makes the "cost" of a 2-vote overstatement 5 times larger than the "cost" of a 1-vote overstatement, where "cost" means the number of additional ballots that must be audited to attain the risk limit. Any value of $\gamma \geq 1$ gives a risk-limiting audit, but γ must be chosen before any ballots have been audited.

3.2   Ballot-polling RLA using BRAVO

The BRAVO ballot-polling RLA by Lindeman, Stark, and Yates (2012) can be applied immediately to constituencies in India. In the Indian scenario, we have only one winner per constituency and one candidate per ballot. For each loser ℓ, the null hypothesis $H_{0w\ell}$ states

---

[4]Indian EVMs (as far as we know) do not produce blank votes. However, if they did they could be accommodated easily. A 1-vote overstatement occurs if we find a blank vote in the reported winner's pile. A netural error would occur when there were at least three candidates and we found a blank vote in a reported loser's pile. A one-vote understatement would occur when there were exactly two candidates and we found a blank vote in the reported loser's pile.

that w did not get more votes than ℓ, that is, that the true outcome was a tie or that ℓ actually won. BRAVO uses Wald's Sequential Probability Ratio Test Wald (1945) to test all the null hypotheses simultaneously.[5]

The audit begins by choosing the risk limit α, the maximum probability of stopping short of a full manual tally if a full manual tally would show a different electoral outcome. It also requires the reported vote totals for each candidate,[6] but no other data from the voting system.

For every apparent loser ℓ, define the *conditional vote share* $s_{w\ell}$:

$$s_{w\ell} \equiv \frac{v_w}{v_w + v_\ell} \tag{3.2}$$

Here, v_w and v_ℓ are the reported vote totals for the winner w and the loser ℓ respectively. If the reported vote tally is correct, the chance that a randomly selected ballot shows a vote for w, given that it shows a vote for either w or ℓ, is $s_{w\ell}$.

BRAVO maintains a test statistic $T_{w\ell}$ for each reported (winner, loser) pair. In Indian elections, there is only one reported winner w per constituency, so this amounts to a test statistic for each reported loser ℓ. Null hypothesis $H_{0w\ell}$ is rejected if:

$$T_{w\ell} \geqslant \frac{1}{\alpha}. \tag{3.3}$$

If the null hypotheses for all apparent losers $\ell \in L$ are rejected, we stop the audit and declare the announced outcome correct.

At any time, for example if the audit is expected to take more time than simply counting the ballots, we can stop the audit and perform a full manual recount. The algorithm runs as follows:

---

[6]There are other ballot-polling methods that do not use the reported results at all.

Algorithm 1 BRAVO with protection against manifest errors.

This is a simplified version of BRAVO that assumes the contest has only one winner, and that there can be at most at most one valid vote per ballot. It incorporates the "phantom to zombie" method of Bañuelos and Stark (2012) for dealing with errors in the ballot manifest.

Input: Risk Limit α; ballot manifest, announced winner w, losers L and corresponding weighted vote shares $s_{w\ell}$ for each $\ell \in L$. Upper bound N on the number of ballots, where N is at least as large as the number of ballots listed in the manifest. Work threshold $K \leq N$.

1: Simultaneous probability ratio: $\forall \ell \in L : T_{w\ell} \leftarrow 1$
2: Number of audited ballots: $n \leftarrow 0$
3: **while** not every $H_{0w\ell}$ has been rejected **do**
4:     Generate a random number i between 1 and N
5:     Look up the ith ballot in the ballot manifest and (attempt to) retrieve it
6:     **if** Ballot i is not in the ballot manifest or cannot be found **then**
7:         For every $\ell \in L$, multiply $T_{w\ell}$ by $2(1 - s_{w\ell})$
8:     **else if** Ballot i shows a vote for the winner w **then**
9:         For each $\ell \in L$, multiply $T_{w\ell}$ by $2s_{w\ell}$
10:     **else if** Ballot i shows a vote for loser $\ell \in L$ **then**
11:         Multiply $T_{w\ell}$ by $2(1 - s_{w\ell})$
12:     **if** $T_{w\ell} \geqslant \frac{1}{\alpha}$ **then**
13:         Reject $H_{0w\ell}$. Withdraw the pair (w, l) from the audit.
14:     **if** the number of ballots inspected exceeds K, or optionally at any time **then**
15:         **STOP** the audit and perform a full manual recount.
16: Declare election outcome correct—all null hypotheses have been rejected.

At any stage, $P = \max_{\ell \in L} 1/T_{w\ell}$ is a conservative sequential P-value for the hypothesis that the reported winner w did not actually win the constituency.

3.2.1   Number of votes to be audited

Consider an example of a 3-candidate contest with a single plurality winner. The candidates are Ramith, Shyam and Priya. Their respective shares is recorded in the following table:

Ramith	Shyam	Priya
$20,000$	$30,000$	$50,000$

In this case, the winner is Priya. Let us denote the winner-loser pairs as (p, r) for Priya and Ramith and (p, s) for Priya and Shyam. The weighted vote shares are:

$$s_{pr} = \frac{v_p}{v_p + v_r} = \frac{50}{50 + 20} = 0.714$$

$$s_{ps} = \frac{v_p}{v_p + v_s} = \frac{50}{50 + 30} = 0.625$$

11

We set the risk limit α at 5%. Every time the audit selects a ballot that shows a vote for Priya we multiply T_{pr} by $\frac{0.714}{0.5} = 1.428$ and T_{ps} by $\frac{0.625}{0.5} = 1.25$. Therefore, the minimum sample size n to attain a risk limit $\alpha = 0.05$ satisfies

$$1.428^n \geq 20$$

and

$$1.25^n \geq 20$$

The smallest such n is $n = 14$. Hence, we need to audit at least 14 ballots—if they all show up votes for Priya, BRAVO will confirm the election outcome at risk limit 5%.

If the reported election results were accurate, on average we would see 50% of ballots for Priya, 30% for Shyam and 20% for Ramith. Lindeman, Stark, and Yates (2012) describe how to find the *Average sample number (ASN)*, the expected sample size necessary to reject all the null hypotheses, assuming the reported results are indeed correct. Stark's online ballot-polling tool shows an ASN of 123 for this example.

3.3   Improved methods for single-constituency RLA's

There have been numerous improvements to the efficiency of Risk-Limiting Audits, any of which could easily apply to India's simple electoral system. See for example `https://github.com/pbstark/` and `https://github.com/pbstark/S157F17/blob/master/pSPRTnoReplacement.ipynb`.

The next section explains how to audit the overall parliamentary winner by an efficient combination of single-constituency audits. It requires independent, sequentially valid P-values $\{P_i\}$ for the hypotheses that the reported outcome in constituency i is incorrect. It does not require the P-values to be obtained using the same method. For instance, some constituencies could use ballot polling and others could use transitive ballot-level comparison audits.

4   Auditing the overall parliamentary winner

A party or a coalition needs a majority of the seats in the Lower House of Parliament to form a new government. The total number of seats is 543, so to win, a party or coalition needs at least 272 seats. The audit needs to confirm that the reported winning party or coalition truly won at least 272 seats. (The particular seats the reported winner won is immaterial to whether they won overall.) If party w supposedly won $M \geq 272$ constituencies, then for a different party to have won in fact, the reported outcome must be wrong in at least $m = M - 271$ of the constituencies that w supposedly won. Note that this condition is necessary but not sufficient for the parliamentary outcome to be wrong: if w in fact won some constituencies it was reported to have lost, the outcome could be wrong in m constituencies w supposedly won and yet w could still be the overall winner. However, if the audit provides strong evidence that there is no set of m constituencies w reportedly won for which w did not actually win, w must be the overall winner.

Let W denote the set of constituencies w reportedly won. Then $|W| \geq 272$ and $m = |W| - 271$, where $|W|$ denotes the cardinality of the set W. If there is no set of constituencies $C \subset W$ with $|C| = m$ for which w lost in *every* $c \subset C$, w really won overall.

Let α denote the overall risk limit, and let P_i denote a P-value for the hypothesis that the reported outcome in constituency i is wrong. We suppose that the audits in different constituencies rely on independently selected random samples of ballots, so the P-values $\{P_c\}$ are independent random variables. If the reported outcome in constituency c is incorrect, the probability distribution of P_c is stochastically dominated by a uniform distribution. That is, $\Pr\{P_c \le p\} \le p$ if the reported outcome in constituency c is wrong.

   Fisher's combining function for a set of P-values $\{P_c\}_{c \in C}$ is

$$X^2(C) \equiv -2 \sum_{c \in C} \ln P_c. \tag{4.1}$$

If the P-values $\{P_c\}$ are independent and all the null hypotheses are true, the probability distribution of $X^2(C)$ is stochastically smaller than a chi-square distribution with $2|C|$ degrees of freedom.[7] That is, if the reported outcome in every constituency $c \in C$ is wrong,

$$\Pr\{X^2_{2|C|} \ge \chi^2_{2|C|}(1 - \alpha)\} \le \alpha, \tag{4.2}$$

where $\chi^2_{2|C|}(1 - \alpha)$ is the $1 - \alpha$ quantile of the chi-square distribution with $2|C|$ degrees of freedom.

   The overall strategy for auditing the parliamentary outcome is thus as follows:

1. Select an overall risk limit α for the parliamentary outcome.

2. Definitions:

   - W denotes the constituencies the reported winning party allegedly won
   - $m \equiv |W| - 271$
   - $\mathcal{W}_m$ denotes the set of all subsets of W with cardinality m
   - $\chi^2 \equiv \chi^2_{2m}(1 - \alpha)$ denotes the $1 - \alpha$ quantile of the chi-square distribution with $2m$ degrees of freedom
   - for any set C of constituencies, $X^2(C) \equiv -2 \sum_{c \in C} \ln P_c$, where P_c is the P-value of the hypothesis that the reported winner in constituency c did not really win, based on the audit sample selected from constituency c so far. Before any data are collected from c, $P_c = 1$.
   - For any collection $\mathcal{U}$ of sets of constituencies, define $\mathcal{U}(c) \equiv \{U \in \mathcal{U} : U \ni c\}$, all sets of constituencies in $\mathcal{U}$ that contain c.

3. Initialization: Set $\mathcal{U} = \mathcal{W}_m$. Select an initial sample size n_c for each constituency c, and draw the initial sample. It is permissible, but not advisable, to let $n_c = 0$ in any constituency $c \notin W$.

4. Audit: While $\mathcal{U}$ is not empty:

   (a) If there is any $C \in \mathcal{U}$ for which every $c \in C$ has been fully hand counted and the hand count has shown that the reported winner was incorrect in every $c \in C$, stop and perform a full hand count of the entire election.

---

[7]See, e.g., Ottoboni, Stark, et al. (2018).

(b) For every constituency c that has been fully hand counted, if the hand count confirms the reported outcome in c, $\mathcal{U} \leftarrow \mathcal{U} \setminus \mathcal{U}(c)$

(c) $\mathcal{U} \leftarrow \mathcal{U} \setminus \{C \in \mathcal{U} : X^2(C) \geq \chi^2\}$

(d) Increase n_c in one or more constituencies $c \in \cup_{U \in \mathcal{U}} \cup_{u \in U} u$ and inspect the additional ballots.[8]

5. If the loop terminates with $\mathcal{U} = \emptyset$, the audit has confirmed the parliamentary outcome at risk limit α.

Proof that the algorithm above is an RLA for the parliamentary outcome. We need to show that if the parliamentary outcome is wrong, the chance that the audit stops without a full manual tally of every constituency is at most α. If the parliamentary outcome is wrong, the reported winner is wrong in every $c \in C$ for some $C \in \mathcal{W}_m$. Suppose there is such a C. If the audit leads to hand counting every $c \in C$, step (4a) ensures that there will be a full hand count of the entire election. Therefore, there will be a full hand count unless C is removed from $\mathcal{U}$. There are two places that sets of constituencies can be removed from $\mathcal{U}$: step (4b) and step (4c). Step (4b) cannot remove C from $\mathcal{U}$, because, by assumption, handcounting any $c \in C$ would not confirm the reported outcome in c. Therefore, the chance that C is *not* fully hand counted is at most the chance that step (4c) removes C from $\mathcal{U}$. But, by construction (through Fisher's combining function applied to the independent constituency-level P-values), that chance is not larger than α. If there is more than one $C \in \mathcal{W}_m$ for which every reported outcome is wrong, the audit must erroneously remove *all* of them at step (b). But the chance of erroneously removing all of them cannot be larger than chance of removing any one of them individualy, which is in turn at most α.

5   Conclusion and future work

We have presented an approach to conduct risk-limiting audits of the national outcome of Indian elections by combining audits conducted in different constituencies using independent samples. Within a given constituency, the audit could use ballot polling, or—with an initial step of sorting VVPATs—transitive ballot-level comparisons. The P-values in different constituencies are combined using Fisher's combining function, for a collection of sets of constituencies. The collection is constructed in such a way that for the reported parliamentary outcome to be wrong, the reported outcome must be wrong in *every* constituency in at least one of the sets. If there is strong statistical evidence that there is no set of constituencies in the collection for which every reported outcome is wrong, that confirms the national parliamentary outcome. The multi-level structure of Indian parliamentary elections

---

[8]The rule for increasing sample sizes could be as simple as "increase every n_c by 25%," or it could be designed to minimize the expected total amount of auditing required, for instance, by preferentially increasing the sample size in constituencies with large margins and taking into account differences in auditing methods in different jurisdictions (ballot polling versus transitive ballot-level comparison). All else equal, when the outcome is correct, auditing an additional ballot is expected to decrease the P-value more the larger the true margin is. Similarly, all else equal, auditing an additional ballot in a jurisdiction conducting a transitive ballot-level comparison RLA is expected to decrease the P-value more than auditing an additional ballot in a jurisdiction conducting a ballot-polling RLA.

makes it possible to have high confidence in the overall parliamentary outcome without necessarily auditing every constituency to a low risk limit. In future research we will address how to schedule increases in sample sizes in different constituencies to minimize the total expected workload, taking into account the reported margins in different constituencies and the auditing methods used in different constituencies.

Acknowledgments

References

Bañuelos, J.H. and P.B. Stark (2012). "Limiting risk by turning manifest phantoms into evil zombies". In: *arXiv preprint arXiv:1207.3413*. https://arxiv.org/abs/1207.3413.

Benaloh, J. et al. (2011). "SOBA: Secrecy-preserving Observable Ballot-level Audit." In: *EVT/WOTE* 11. http://static.usenix.org/events/evtwote11/tech/final_files/Benaloh.pdf.

Calandrino, J.A., J.A. Halderman, and E.W. Felten (2007). "Machine-Assisted Election Auditing." In: *EVT* 7. http://static.usenix.org/event/evt07/tech/full_papers/calandrino/cala pp. 9–9.

Hall, J.L. et al. (2009). "Implementing Risk-Limiting Audits in California". In: *EVT/WOTE*. URL: http://arxiv.org/abs/0905.4691.

Kroll, A. Halderman, and E. Felten (2014). "Efficiently Auditing Multi-Level Elections". In: URL: https://www.jkroll.com/papers/evote14_multilevel_audit.pdf.

Lindeman, M. and P.B. Stark (2012). "A Gentle Introduction to Risk-limiting Audits". In: URL: https://www.stat.berkeley.edu/~stark/Preprints/gentle12.pdf.

Lindeman, M., P.B. Stark, and V. Yates (2012). "BRAVO: Ballot-polling Risk-limiting Audits to Verify Outcomes". In: URL: https://www.usenix.org/conference/evtwote12/workshop-program/

Ottoboni, K., M. Bernhard, et al. (2019). "Bernoulli Ballot Polling: A Manifest Improvement for Risk- Limiting Audits". In: http://arxiv.org/abs/1812.06361. Voting '19.

Ottoboni, K., P.B. Stark, et al. (2018). "Risk-Limiting Audits by Stratified Union-Intersection Tests of Elections (SUITE)". In: *International Joint Conference on Electronic Voting*. Full version at https://github.com/pbstark/CORLA18/blob/master/CO18.pdf. Springer, pp. 174–188.

Schürmann, C. (2016). "A Risk-Limiting Audit in Denmark: A Pilot". In: Springer. DOI: 10.1007/978-3-319-52240-1_12. URL: https://link.springer.com/chapter/10.1007/978-3-319-

Stark, P.B. (2008). "Conservative statistical post-election audits". In: 2(2), pp. 550–581. DOI: 10.1214/08-AOAS161. URL: https://arxiv.org/abs/0807.4005.

Stark, P.B. (2009). "Risk-limiting post-election audits: *P*-values from common probability inequalities". In: *IEEE Transactions on Information Forensics and Security* 4, pp. 1005–1014.

Stark, P.B. (2010). "Super-Simple Simultaneous Single-Ballot Risk-Limiting Audits". In: USENIX Association Berkeley, CA, USA ©2010. URL: https://www.usenix.org/legacy/event/evtwo

Stark, P.B. (2018). *An Introduction to Risk-Limiting Audits and Evidence-Based Elections: Testimony to the Little Hoover Commission.* https://www.stat.berkeley.edu/~stark/Preprints/lh

Stark, P.B. and V. Teague (2014). "Verifiable european elections: risk-limiting audits for D'Hondt and its relatives". In: *USENIX Journal of Election Technology and Systems (JETS)* 1(3), pp. 18–39.

Stark, P.B. and D. Wagner (2012). "Evidence-based elections". In: *IEEE Security & Privacy* 10(5). https://www.stat.berkeley.edu/~stark/Preprints/evidenceVote12.pdf, pp. 33–41.

Supreme Court of India (2013). *CIVIL APPEAL NO.9093 OF 2013—Dr. Subramanian Swamy vs. Election Commission of India.* http://sci.gov.in/jonew/judis/40874.pdf.

Wald, A. (1945). "Sequential Tests of Statistical Hypotheses". In: *Annals of Applied Statistics* 16(2), pp. 117–186. DOI: 10.1214/aoms/1177731118. URL: http://projecteuclid.org/euclid.aoms/

Wolchok, S. et al. (2010). "Security analysis of India's electronic voting machines". In: *Proceedings of the 17th ACM conference on Computer and communications security.* ACM, pp. 1–14.